MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2018/2019

TDS2101 – INTRODUCTION TO DATA SCIENCE

(All sections / Groups)

12 MARCH 2019
2.30 p.m. – 4.30 p.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 5 pages with 4 Questions only, excluding the cover page.

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the Answer Booklet provided.

Question 1

a) Big Data requires new ways of handling, storing and processing data. Contrast between the **structure of data** in more traditional data analytics and those found in modern Big Data analytics.

(2 marks)

b) There are six dimensions of Data Quality that must be adhered to when dealing with Big Data. Two of the six dimensions are *timeliness* and *validity*.

i. Describe the *timeliness* quality of data.

ii. Explain TWO possible issues that may arise when trying to ensure *validity* of phone numbers in Malaysia.

(3 marks)

c) For a data scientist, business skills are as important as technical skills. Discuss TWO of such business skills that would be crucial to become an accomplished data scientist.

(2 marks)

d) A data scientist is given patients' data containing their dietary consumption (carbohydrates, protein, fibre, etc.) and lifestyle attributes (steps, inactive time), and whether they are diabetic or not. There are a number of questions that he/she should formulate first before attempting to deal with the data.

i. If the data scientist wants to know the average number of steps per day each patient takes, what *type of question* (among the known six types) should he/she formulate?

ii. If the data scientist wants to study the relationship between the consumption of carbohydrates and diabetes, what *type of question* (among the known six types) should he/she formulate?

iii. Provide an example of a **Predictive** question based on the same scenario.

(3 marks)

Continued...

Question 2

The following dataset collected from a new fitness tracker Fitology shows the summary of a user's daily data, including the number of steps, amount of calories burned, and the inactive time in hours and minutes. In addition, the dataset also compiles an optionally recorded data (through Fitology's app) indicating if the user went for a hiking activity.

	Tracker data			User recorded data
date	Steps	Calories Burned	Inactive Time	Hiking
2018-05-13	3839	251	6h 52m	N
2018-05-14	4703	462	6h 35m	Y
2018-05-15	741	65	7h 33m	N
2018-05-16	3143	225	7h 19m	N
2018-05-17	3534	245	7h 10m	N
2018-05-18	4013	255	7h 12m	N
2018-05-19	3325		6h 55m	N
2018-05-20	6544	514	7h 15m	
2018-05-21	3291	235	6h 42m	

a) Identify any outliers based on the *Steps* attribute (indicate the date index where it occurs). You are required to show how you arrive at your answer.

(3 marks)

b) Identify any errors/inconsistencies in the *Calories Burned* attribute (indicate the date index where it occurs). Discuss ONE suitable approach that can be adopted to handle this problem.

(2 marks)

c) If you intend to use values in the *Inactive Time* attribute for further numerical analysis, suggest a way to prepare this data.

(1 mark)

d) Explain in brief steps **how** a model can be trained to classify if the user has gone for a hiking trip on 2018-05-20 and 2018-05-21.

(2 marks)

e) By collecting such data (as shown in the sample dataset) from 900 active users of the Fitology tracker, suggest a technique that you can employ to help the company profile their users into a number of distinct lifestyles and give ONE suggestion on how the company can leverage such insights for their business.

(2 marks)

Continued...

Question 3

a) The CEO of a new online grocer FreshToGo is planning to launch a sales campaign on its mobile application by broadcasting a basket of discounted items in attempt to boost sales of under-performing brands. As the data scientist of FreshToGo, you are to help the CEO to determine *which combination of items*, and of *which brands*, should be broadcasted together at a discounted price to certain users. Identify the steps needed to perform this analysis and for each step, state the choice of models/algorithms (if any) that you may need for this purpose.

(4 marks)

b) Define the meaning of a **confounding variable**, and support your definition with an example.

(3 marks)

c) One important aspect in a data scientist's professional code of conduct that relates to client relationship is **confidentiality of information**. Provide TWO practices that can safeguard inadvertent or unauthorized disclosure or access to data/information belonging to the client.

(3 marks)

Continued...

Question 4

a) Write a line of Python code to perform the following operations on a DataFrame `df`, consisting of `country` (in string) as index, and 3 columns of data named: `capital, population, gdp`, which respectively contain the names of capital cities (in string), population (in integer) and GDP (in integer). Assume that there is no missing data. You do not need to write a full program.

i. Retrieve the 3rd row of the DataFrame.
ii. Compute the mean population of all countries in the DataFrame.
iii. Normalize the GDP data by standardization (i.e. normalize by subtracting the mean, and dividing by its standard deviation).

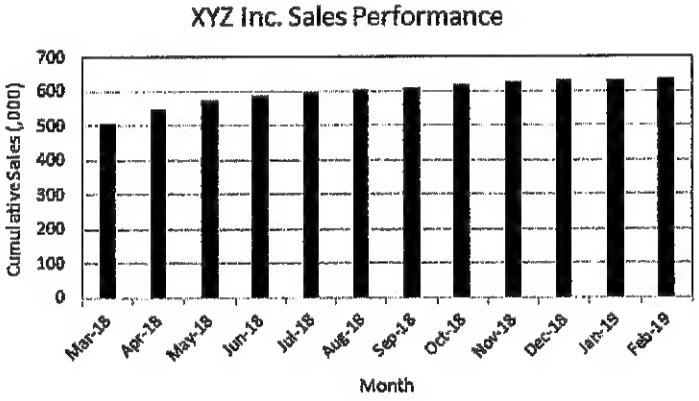
(3 marks)

b) Suggest ONE reason as to why R may be preferred over other languages for Data Science projects.

(2 marks)

c) Monthly sales for XYZ Inc. has been declining for over about a year now. Discuss why the following bar chart is both misleading and also not an appropriate type of visualization. Elaborate on the chart type and choice of data.

(3 marks)



d) What is a **choropleth map**? What is ONE important guideline to ensure it provides meaningful patterns and relationship between different areas/regions?

(2 marks)

End of Paper